北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.039

英汉机器音译系统对比研究

高恩婷[1] 段湘煜[2,†]

1. 苏州科技大学电子与信息工程学院, 苏州 215009; 2. 苏州大学计算机科学与技术学院, 苏州 215006;
† 通信作者, E-mail: xiangyuduan@suda.edu.cn

摘要 针对机器音译的两种主要方法——传统的基于统计的方法和目前流行的基于深度神经网络的方法, 分别使用两种典型系统进行研究。实验结果显示, 基于统计的方法和基于深度神经网络的方法取得的音译质量在评测指标上相当, 但在具体音译结果上各系统间呈现不一致的输出。使用系统融合的方法来实现各系统间的优势互补。实验结果显示, 系统融合的方法显著优于单系统的音译质量。

关键词 机器音译; 音译对齐; 统计方法; 深度神经网络方法

中图分类号 TP391

A Comparative Study on English-Chinese Machine Transliteration

GAO Enting[1], DUAN Xiangyu[2,†]

1. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009;
2. School of Computer Science and Technology, Soochow University, Suzhou 215006;
† Corresponding author, E-mail: xiangyuduan@suda.edu.cn

Abstract With the aim to study the two main methods on machine transliteration: traditional statistical method and the current prevalent deep neural network method, the authors carry out the comparative study on them with two typical systems per method The experiments show that traditional statistical method and deep neural network method perform comparatively regarding evaluation metrics, while manifest difference on individual transliteration result. A system combination method is proposed to balance the strengths of all systems. Experimental results show that system combination significantly improves the transliteration quality over single system.

Key words machine transliteration; transliteration alignment; statistical method; deep neural network method

音译指将一种语言的单词或短语按照发音规则翻译成另外一种语言。这里的“单词或短语”一般指人名、地名、组织结构名和专业术语等。比如英文名字“Smith”被音译为汉语的“史密斯”, 其音译过程要保持发音一致或相似, 但用另外一种语言的文字进行书写。机器音译借助计算机进行自动音译, 可以用于语料对齐、跨语言信息检索和抽取, 也可以为机器翻译(MT)任务中 OOV (out-of-vocabulary)的处理提供有益的补充。

传统的机器音译方法主要基于统计模型来完成音译过程[1−2]。随着深度学习的方法在自然语言序列任务中越来越多的成功应用, 深度神经网络也在机器音译领域受到关注[3]。目前, 还没有关于这两种方法在机器音译领域的比较研究。本文对传统的统计机器音译方法和目前流行的深度神经网络的方法进行对比, 以便发现两种方法各自的优缺点, 进一步提升机器音译系统的性能。针对传统的统计机器音译方法, 我们尝试两个典型的系统: 联合信道模型系统[1]和噪声信道模型系统[4]; 针对深度神经网络机器音译方法, 我们构建一个基于目标端双向 LSTM 协议网络的音译系统以及一个基于注意力机制的神经网络音译系统。

国家自然科学基金(61273319, 61373095)资助
收稿日期: 2016−07−29; 修回日期: 2016−10−06; 网络出版日期: 2016−11−28

实验结果显示, 基于统计的机器音译方法和基于神经网络的机器音译方法取得相当好的音译性能, 但两个方法在具体音译输出上不尽相同。本文使用基于混淆网络的系统融合方法, 对这两个方法的输出进行融合, 进一步提升了机器音译的性能。

1 统计机器音译方法

统计机器音译方法是通过对平行音译语料库 $\{E, C\}$进行统计分析, 构建统计音译模型, 其中 E 表示源语言端的语料, C 表示目标端相应的音译语料。根据统计模型的不同, 基本上分为以下两类典型的模型: 联合源信道模型(Joint Source Chanel Model, JSCM)[1]和噪声信道模型(Noise-Chanel Model, NCM)[4]。

1.1 联合源信道模型(Joint Source Chanel Model)

受 N-gram 语言模型的启发, 联合源信道模型将某一音译对 E 和 C 分解为各个子音译单元, 则联合概率 $P(E, C)$被分解为 N-gram 子音译单元的概率乘积:

$$\begin{aligned} P(E,C) &= P(e_1e_2...e_k, c_1c_2...c_k) \\ &= \prod_{k=1}^{K} P(<e,c>_k | <e,c>_1^{k-1}) \\ &\approx \prod_{k=1}^{K} P(<e,c>_k | <e,c>_{k-n+1}^{k-1}) 。 \end{aligned}$$

E 和 C 被分解为 k 个子音译单元, 例如“SMITH”和“史密斯”可以分解为 3 个子音译单元: <S, 史>、<MI, 密>和<TH, 斯>。此联合概率在马尔科夫假设的前提下, 各个子音译单元的条件概率仅以前 $n-1$ 个子音译单元为条件(子音译单元的分解方法将在 1.3 节介绍)。N-gram 子音译单元语言模型的训练使用 SRILM 工具包[5], 数据平滑使用 Kneser-Ney 平滑方法[6], 在音译解码过程中, 我们使用基于 beam search 的网格解码(lattice-decoding)来生成 n-best 音译结果[5]。

1.2 噪声信道模型(Noise-Chanel Model)

受统计机器翻译[4]的启发, 在传统的统计机器音译方法中, 噪声信道模型被广泛应用。给定一个英语姓名 E, 该模型将寻找相应的汉语音译结果 C 以最大化概率 $P(C|E)$。根据贝叶斯理论, 意味着要找到一个 C 使得后验概率最大化:

$$\begin{aligned} P(C|E) &\propto P(E|C) \times P(C) \\ &= \prod_{k=1}^{K} P(e_k|c_k) \times P_{\mathrm{LM}}(C)^{\lambda_{\mathrm{LM}}} \\ &= \prod_{k=1}^{K} \prod_{i=1}^{I} \varphi_i(e_k, c_k)^{\lambda_i} \times P_{\mathrm{LM}}(C)^{\lambda_{\mathrm{LM}}} 。 \end{aligned}$$

我们采用对数线性模型, 将 $P(E|C) \times P(C)$的计算分解为各个特征之积。其中, $P_{\mathrm{LM}}(C)$为语言模型概率特征, 可以通过目标端基于字符的 N-gram 语言模型计算而得, 其权重为 λ_{LM}; $P(E|C)$为音译模型概率, 其因式分解为各个子音译单元的条件概率之积, 各个子音译单元的条件概率进而被分解为 I 个特征 Φ, 每个特征附有一个权重 λ。特征包括: 正向子音译单元概率之积以及反向子音译单元概率之积; 正向字符对齐权重以及反向字符对齐权重; 字符个数惩罚因子; 子音译单元个数惩罚因子。

由于在机器音译任务中字符对齐是单序(monotone)的, 没有调序, 因此我们采用 1.3 节介绍的单序的音译对齐方法, 而没有采用对于调序没有限制的 GIZA++对齐工具[7]。解码器使用基于短语的机器翻译工具 Moses[8]。

1.3 音译对齐(Transliteration Alignment)

上述两个音译系统均以音译对齐为基础, 在获得音译对齐结果后训练音译模型的解码器。音译对齐指在机器音译过程中寻找等价发音单元的过程, 比如在音译对“SMITH|史密斯”中, 等价发音单元依次为“S|史”、“MI|密”、“TH|斯”。以发音方式作为中枢, 可以将汉语的表面字符和英语发音单元建立联系。由于汉语字符不带有发音方式信息, 我们采用将汉语字符先转化为拼音, 再由拼音转化为发音音素(phoneme)的方法获得汉语的发音方式。汉语的发音音素采用 IIR 音素集合[9]。如图 1 所示, 通过将汉语名字转换为音素序列, 便可以利用发音方式作为中枢, 寻找相应的英语音节, 同时由于汉语字符通常映射到一个唯一的音素序列, 故汉语音素与英文音节的对齐可以直接转化为汉字与英语发音单元的对齐。

由于没有等价发音单元的人工标注, 我们使用无监督产生的方法, 在给定音译对语料的条件下, 首先将所有汉字转化为音素序列, 并重新实现基于音译距离的 EM 算法[2], 以获得汉语音素与英语发音单元的对齐(即图 1 中目标端音素序列与源端的对齐), 最后再将汉语音素转换为汉字。

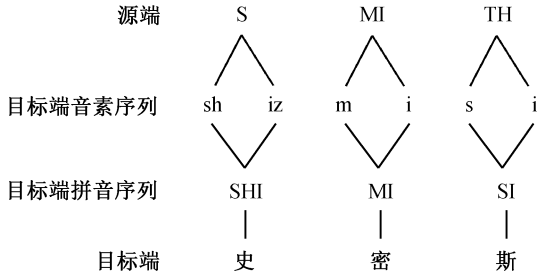


图 1　音译对齐示例
Fig. 1　Transliteration alignment example

2　神经网络机器音译方法

根据是否使用音译对齐，本文尝试两种神经网络系统：基于目标端双向 LSTM 协议网络的音译系统，标记为“Agreement”；基于注意力机制的神经网络音译系统，标记为“Attention”。Agreement 系统不尝试对音译对齐进行建模，只对序列–序列进行建模；Attention 系统在音译过程中使用注意力机制，对序列–序列任务中隐含的音译对齐结构进行建模。两种方法的系统框图见图 2。

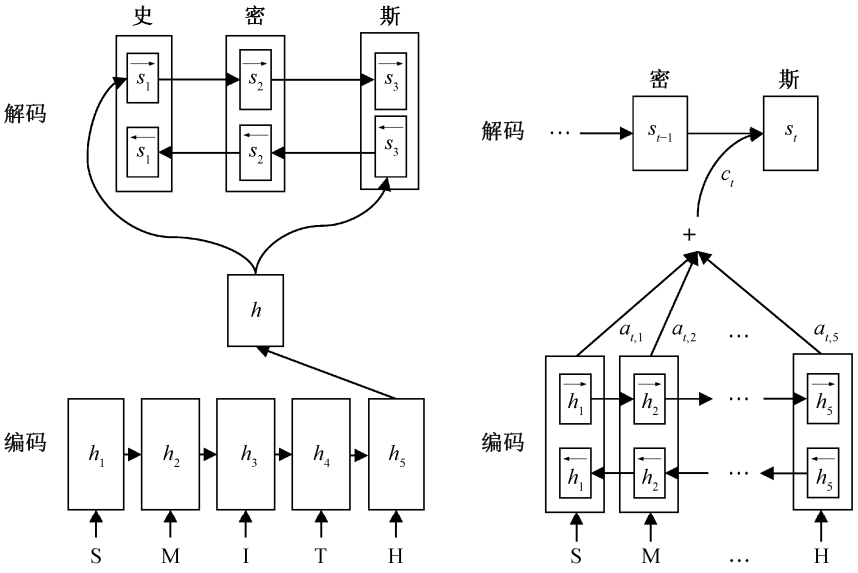
2.1　Agreement 系统

传统循环神经网络系统虽然已经成功地运用于语言模型等分析任务中，但存在明显的弱点。如在序列–序列任务中，由于解码时需要根据上下文向量和前一时刻的输出做判断，故前一时刻的输出错误会影响当前的判断，造成错误累积。尤其在长序列的任务中，该问题更加明显，使得输出中的前缀质量较高而后缀质量较低。Agreement 系统使用双向输出，可以减弱传统的序列–序列方法中错误容易累加的缺点，如图 2(a)所示。该方法使用成对的长短期记忆循环神经网络(LSTM RNN)，首先将源端序列编码成一个固定长度的向量，分别从左向右和从右向左两个方向进行解码输出，在两个方向上分别产生 *k*-best 输出，并设计一个机制来鼓励这两个 *k*-best 列表相互间的一致性(agreement)。Liu 等[10]的实验显示，该双向网络可以产生更加平衡的输出，但不需要对音译对齐进行建模处理。

2.2　Attention 系统

在神经网络机器翻译研究中，编码–解码结构(Encoding-Decoding)被广泛采用。随着注意力机制的引入，神经网络机器翻译系统获得大幅度的性能提升[11]。我们在机器音译任务上应用带有注意力机制的编码–解码结构，如图 2(b)所示。

由于未使用音译对齐，源端的英语子音译单元并没有分割出来。在各个源端字符之上，构建一个双向的循环神经网络，以准确捕捉源端序列上各点



(a) Agreement系统　　(b) Attention系统

s 和 *h* 分别为源端和目标端词的隐含向量，向右和向左箭头分别表示从左向右计算和从右向左计算

图 2　神经网络机器音译系统
Fig. 2　Neural network transliteration systems

的局部信息。在双向循环神经网络之上, 引入注意力机制(即目标端的当前时刻 t 与所有源端各个 h 向量的关系), 以获得目标端与源端的软性对齐(soft alignment), 其输出 c_t 是目标端与源端的期望对齐, 与目标端上一时刻 s_{t-1} 构成解码层的循环神经网络, 逐次生成目标端的音译结果。

注意力机制是在训练过程中逐渐计算得出软性音译对齐, 而不是像第 1 节中的系统均在训练之前获得子音译单元的对齐。与 Agreement 系统相比, 该系统除增加注意力机制外, 还包括其双向 RNN 是在源端进行编码, 而 Agreement 的双向 RNN 是在目标端进行解码。

3 系统融合

我们使用混淆网络(Confusion Network)将上述 4 个系统的音译结果进行融合, 系统融合步骤[12]如下。1) 主干选择(Backbone Selection): 从所有系统的音译输出中选择一个主干, 用于决定最终融合结果的字符顺序。2) 音译输出的字符对齐: 用于在主干和所有系统的音译输出间构建字符对齐。3) 混淆网络的构建: 在第 2 步构建的音译输出对齐的基础上构建混淆网络。4) 混淆网络解码: 在混淆网络中寻找最优的音译路径。在上述 4 个步骤中, 由于不同的音译系统输出的字符不同, 字符出现的顺序也不同, 使得第 2 步的音译输出对齐最具挑战性。我们使用以下 4 种词对齐方法, 进行音译输出的字符对齐。

1) GIZA++[7]: 使用由 IBM 模型 I 自举得出的 HMM 模型, 获得主干和所有系统的音译输出间的字符对齐。收集测试集所有输入的所有音译输出(各系统的音译输出), 以构成 GIZA++的训练集合, 其间允许多对一的字符对齐。

2) TER[13]: 即翻译错误率, 用于测度从音译结果到答案(在系统融合应用中即主干)所需的最少编辑操作步数。编辑步骤包括插入、删除、替换和音节迁移。最优对齐就是使得编辑步骤最少的字符对齐, 其间只允许一对一的字符对齐。

3) CLA[14]: 即竞争连接算法。运用贪婪算法, 搜索主干和各系统音译输出间具有最高相关度的字符对齐。CLA 算法只输出一对一的字符对齐。

4) IHMM[15]: 即非直接隐马尔科夫模型, 模型参数由各种特征非直接地计算得出, 特征包括语义相似度、扭曲惩罚因子等。IHMM 输出多对一的字符对齐。

4 实验

4.1 语料

本文利用新华网提供的《外国人名音译词典》(Chinese Transliteration of Foreign Personal Names)作为机器音译的训练和测试语料。该词典是当代汉语出版对于外国人名进行音译的标准, 包括 37694 个不重复的英文姓名及其相应的官方汉语音译, 其中不同的英文姓名可能对应同一个汉语姓名, 如“DENNIN”和“DENNING”均对应“丹宁”。这些英文姓名来自英语国家、法语国家、西班牙语国家、德语国家、阿拉伯语国家、俄语国家以及其他语言国家。本文在英汉音译和汉英音译两个任务上进行实验, 表 1 列出数据划分的详细信息。

4.2 评测指标

机器音译质量由 4 个评价指标进行评测, 对每个系统的 n-best 输出中, 我们选择 10-best 结果进行评测。某些源端的姓名可能会有多个正确的目标端姓名。多正确目标端名在我们的评测中认为同等重要, 即正确目标名中的任意一个被匹配都认为正确, 系统的 n-best 输出中第一个匹配正确答案的条目便作为一个正确的条目。

由于存在多答案、多输出(n-best 结果)的复杂情况, 我们使用以下 4 种评测指标①。

1) TOP-1 准确率(ACC): 也叫做词错误率, 用于评测音译 n-best 结果中第一个结果的准确率, ACC=1 代表所有条目的第一个音译结果都正确, ACC=0 则代表都错误。

2) 平均 F 值(Mean F): 用于评测音译 n-best 结果中第一个结果与最相近答案的区别度, 通过二者之间公共子串的长度进行估计。Mean F=1 代表所有的第一个结果都能匹配答案, Mean F=0 代表所

表 1 音译数据

Table 1 Transliteration data

源端	目标端	训练集/K	开发集/K	测试集/K
英语	汉语	37	2.8	1.0
汉语	英语	28	2.7	2.2

① 评测脚本来自 NEWS 2015: http://www.colips.org/workshop/news2015/index.html

有的第一个结果与所有答案都没有公共子串。

3) 平均互惠级(Mean Reciprocal Rank, MRR): 用于评测音译结果的平均正确级。MRR 越接近 1, 代表音译结果的 *n*-best 中的高级别条目越接近正确答案。

4) 最大后验(MAP_{ref}): 用于评测音译 *n*-best 结果中正确答案的百分比。

4.3 实验配置

本文统计机器音译方法中均使用音译对齐, 重新实现基于音素对齐的音译对齐工具[2], 同时也在字符级别的对齐任务上尝试了 GIZA++。在联合源信道模型和噪声信道模型中, 其 N-gram 模型部分均使用 SRILM 工具包[5], 两个模型都输出 20-best 结果用于评价音译质量。

神经网络机器音译方法中, Agreement 系统采用 Agtarbidir 工具包①, Attention 系统采用 Block 工具包②。两者均经过 100 epochs 的训练过程, 所有的训练过程都在一个 Tesla K40m GPU 上完成。两者在源端和目标端的 RNN 均使用 500 维的嵌入向量, minibatch 的大小为 16, beam search 的宽度设为 12, 用于 *n*-best 输出。使用 AdaDelta 学习率[16]进行迭代, 梯度的 clipping threshold 设为 1, dropout rate 设为 0.6, 两个系统在源端和目标端的 vocabulary size 均设为 500, 长度超过 20 的音译对不作为训练数据。除此之外, Agreement 系统使用多个神经网络融合的方法, 在进行 beam search 时, 对每个目标端词的概率使用简单的线性插值(每个神经网络具有相同的权重)进行融合, 选出每 5 轮的双向 RNN 进行线性插值。

4.4 实验结果与分析

4.4.1 英–汉音译

本文探讨的 4 个系统的音译质量评测结果如表 2 所示, 其中 JSCM 和 NCM 为统计机器音译方法, Agreement 和 Attention 为神经网络机器音译方法。从表 2 可以看出, 统计机器音译方法和神经网络机器音译方法都取得相近的音译质量。就评价指标 ACC 和 MAP_{ref}而言, JSCM 的性能最好; 就评价指标 Mean *F* 和 MRR 而言, Agreement 方法取得最好的性能。

基于编码–解码的神经网络机器音译系统并非都能取得与统计机器音译系统相当的性能, Agreement 方法显著优于 Attention 方法。Agreement 在解码阶段进行从左到右和从右到左的双向解码, 并在 beam search 过程中寻找双向的一致性, 克服了传统的编码–解码系统易产生较好的输出前缀, 而后缀往往错误, 从而引起整体音译结果的准确率较低的缺点。实验显示, Agreement 系统可以生成高质量的整体音译输出。Attention 系统虽然引入了具有软性音译对齐功能的注意力机制, 并在机器翻译任务上取得显著成功[11], 但在机器音译任务上, 其性能显著低于不具有音译对齐功能的 Agreement 系统, 说明在神经网络机器音译任务中音译对齐的使用并不能提升系统性能。在统计机器音译任务上, 音译对齐的使用则显示出有效性, 如表 3 所示。

表 2　各个系统的英–汉音译输出的评测结果

Table 2　English-Chinese transliteration evaluations on all systems

系统	ACC	Mean *F*	MRR	MAP_{ref}
JSCM	0.3194	0.6588	0.3973	0.3089
NCM	0.3144	0.6598	0.3906	0.2985
Agreement	0.3165	0.6643	0.4130	0.3086
Attention	0.2539	0.6177	0.3339	0.2408

针对两种统计机器音译系统, 我们比较了两个方面的实验效果: 1) N-gram 模型阶数的影响; 2) 音译对齐工具的影响。表 3 列出对这两个方面进行实验的结果。

就 N-gram 模型的阶数而言, 由于音译数据往往长度较短, 所以我们测试了 2-gram 和 3-gram。JSCM 和 NCM 系统都呈现相同的效果, 2-gram 的性能显著优于 3-gram。就音译对齐工具而言, 我们在 NCM 系统上测试了两个对齐工具: 基于音素对齐的音译对齐工具[2]和 GIZA++对齐工具。由于音译过程是一个单向的过程, 为此过程量身定制的音译对齐工具只允许单向对齐, 而 GIZA++却允许任意方向的对齐, 这会引起不合理的子音译单元的对齐。实验结果显示音译对齐工具的效果优于 GIZA++的效果, 本文所重现的音译对齐工具与音译过程更为契合。

表 3 中, 联合源信道模型 JSCM 显著优于噪声信道模型 NCM, 表明将音译过程建模为源端–目标

① https://github.com/lemaoliu/Agtarbidir

② https://github.com/mila-udem/blocks

表 3 统计机器音译方法在 N-gram 和对齐方法上的性能比较

Table 3 Comparisons on N-grams and alignment methods between statistical machine transliteration systems

系统	对齐工具	N-gram	ACC	Mean *F*	MRR	MAP_{ref}
JSCM	音译对齐	2-gram	0.3194	0.6588	0.3973	0.3089
		3-gram	0.3125	0.6587	0.3969	0.3044
NCM	音译对齐	2-gram	0.3144	0.6598	0.3906	0.2985
		3-gram	0.3035	0.6537	0.3801	0.2880
	GIZA++对齐	2-gram	0.3115	0.6589	0.3896	0.2962
		3-gram	0.3009	0.6541	0.3763	0.2860

端联合的 N-gram 模型更为有效。

4 个系统中, JSCM, NCM 和 Agreement 的性能指标较为接近, 而 Attention 的性能显著低于其他系统。对于性能指标接近的 3 个系统, 虽然呈现相近的评测结果, 但 3 个系统的输出却并不相近。表 4 比较各个系统的 TOP-1 结果, 与评测指标中的 ACC 相对应, 该混淆矩阵计算 4 个系统间两两相一致的比率。

从表 4 可看出, 4 个系统间输出最为一致的是 JSCM 和 NCM, 两者均为统计机器音译方法, 并且均为基于音译自动对齐的结果。虽然这两个系统最为一致, 但具体的比率却为 0.7232, 相对较低, 表明两个系统的输出还存在很多不一致的结果。4 个系统之间最不一致的是 NCM 与 Attention, 只有 0.5089 的结果是一致的, 主要是由于 Attention 的性能比其他 3 个系统都明显低。使用神经网络方法的 Agreement 和 Attention 系统一致性最高, 与使用统计方法的 JSCM 和 NCM 系统的一致性都很低。

表 4 中系统间两两相一致的比率从 0.5089 至 0.7232 不等, 其一致性的比率总体上较低, 表明各个系统有自身的优势, 因此可以在系统与系统间构建一个机制, 以利用系统互补的特质。使用基于混淆网络的系统融合方法后取得的性能结果如表 5 所示, 可以看到系统融合方法显著提高了单系统的音译质量, 其中基于 GIZA++对齐的融合方法获得的效果最佳。

表 4 ACC 混淆矩阵

Table 4 Confusion matrix on ACC

系统	NCM	Agreement	Attention
JSCM	0.7232	0.6517	0.5357
NCM		0.6011	0.5089
Agreement			0.5763

表 5 系统融合结果

Table 5 System combination results

对齐方式	ACC	Mean *F*	MRR	MAP_{ref}
GIZA++	0.3306	0.6693	0.4191	0.3146
TER	0.3217	0.6643	0.4103	0.3103
CLA	0.3210	0.6619	0.4069	0.3093
IHMM	0.3269	0.6665	0.4159	0.3109

为了直观地反映系统差异, 我们将各系统在测试集的一个片段(包含 6 个英文姓名)上的音译结果列于表 6, 其中统计方法(JSCM 和 NCM)都以音节对齐为基础, 而神经网络机器音译方法(Agreement 和 Attention)虽然不以音节对齐为基础, 但其 end-to-end 的音译结果仍能体现音节上的对应关系。就与答案的匹配程度而言, Attention 方法的匹配程度较低。

4.4.2 汉–英音译

表 7 列出在汉–英音译任务上各个方法的性能。可以看出, 统计机器音译方法中 JSCM 的性能略好于 NCM, 神经网络机器音译方法中 Agreement 方法显著优于 Attention 的方法。同机器音译在英–汉任务上的表现相似, 汉–英任务中 JSCM 和 Agreement 的性能效果相近, 表明统计方法和神经网络方法可以取得相近的音译质量。

汉–英音译任务是英–汉音译任务的反向音译(back transliteration)任务, 通常比其正向任务困难。对比表 7 和表 2 可看出, 汉–英任务的表现弱于英–汉任务, 而表 7 中的 Mean *F* 高于表 2 中的 Mean *F*。这是由于 Mean *F* 是测量音译结果与答案之间的公共子串, 而英语作为目标端, 其单个英文字母很容易与答案的子串匹配。

表 6　英–汉音译实例

Table 6　English-Chinese transliteration example results

英文	答案	JSCM	NCM	Agreement	Attention
AALTONEN	阿尔托宁	阿尔托宁	阿尔托嫩	阿尔托宁	亚尔托宁
ABANO	阿巴诺	阿巴诺	阿巴诺	阿巴诺	阿巴诺
ABRA	阿布拉	阿布拉	阿布拉	阿布拉	艾布拉
ACHILLE	阿希莱	阿基利	阿基利	阿奇利	阿基尔
ADINA	阿迪纳	阿迪纳	阿迪娜	阿迪纳	阿迪纳
ADJANI	阿贾尼	阿德甲尼	阿德甲尼	阿甲尼	阿迪甲尼

表 7　各个系统的汉–英音译输出的评测结果

Table 7　Chinese-English transliteration evaluations on all systems

系统	ACC	Mean F	MRR	MAP_{ref}
JSCM	0.2161	0.7210	0.2937	0.2093
NCM	0.2041	0.7172	0.2909	0.1996
Agreement	0.2139	0.7292	0.3026	0.2119
Attention	0.1567	0.6679	0.2403	0.1593

5　讨论与展望

Attention 和 Agreement 方法都是针对 Encoder-Decoder 的原始架构在长句上性能急剧下降的缺点提出的，二者从不同角度来克服这个缺点：Attention 方法避免了将源端的句子编码成一个定长的向量(如果将长句和短句都编码成一个定长的向量，会造成长句的某些信息在定长向量中遗失)，而是将源端句子的信息分布在各个源端词上，长句的信息不会遗失；Agreement 方法是解决目标端解码的错误传播问题(句子越长，错误从左向右传播越严重)，通过双向解码并在双向中寻找 Agreement，可以减弱错误传播问题。

由于音译数据往往较短，机器音译错误的产生和传播比较容易捕捉和纠正，Agreement 方法通过双向解码并寻找一致性的音译结果，较易发现错误和纠正错误。但是，Agreement 方法还没有应用到 MT 任务中，主要是因为 MT 的数据较长，错误传播造成双向解码较难寻找到一致性的结果。

Attention 方法在 MT 任务中有成功的应用，但其软性词对齐效果仍然显著落后于传统的词对齐工具(如 GIZA++等)。在汉英 MT 任务上，对齐错误率 AER 的比较结果为 54 对 30，Attention 方法落后传统词对齐工具 24 个点[17]；在机器音译任务上，Attention 方法的音节对齐也会显著弱于基于统计的对齐工具，而音节对齐在机器音译任务中很重要，故 Attention 方法的效果不如其他方法理想。

以上两种神经网络方法是从不同的角度克服长句性能瓶颈问题。在未来的工作中，可以将这两个角度进行结合，即在 Attention 机制上引入双向解码，并寻求一致性的解码结果，可能会提升机器音译和机器翻译的性能。

6　结语

本文分别针对基于统计的机器音译方法和基于神经网络的机器音译方法进行对比研究。基于统计的方法中使用两种典型模型，分别是联合源信道模型和噪声信道模型，二者最根本的区别是分别对音译的联合概率和条件概率进行建模。基于神经网络的方法中也使用两种典型模型，分别是目标端双向 LSTM 协议(Agreement)神经网络和注意力(Attention)神经网络，二者都基于编码–解码(Encoder-Decoder)框架，主要区别是双向解码还是双向编码。虽然 Attention 方法在机器翻译任务上取得成功，但在机器音译任务上与其他方法相比却取得最低的性能。Agreement 方法与本文中使用的两个传统的基于统计的方法性能相当，表明基于神经网络的方法可以取得与传统方法相近的音译质量。本文使用基于混淆网络的系统融合方法对上述 4 个系统进行系统融合，进一步提升了音译输出质量。

参考文献

[1] Li Haizhou, Zhang Min, Su Jian. A joint source-channel model for machine transliteration // Scott D, Daelemans W, Walker M A. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona: ACL, 2004: 159–166

[2] Pervouchine V, Li H, Lin B. Transliteration alignment // Su K Y, Su J, Wiebe J. Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapore: ACL, 2009: 136–144

[3] Finch A, Liu L, Wang X, et al. Neural network transduction models in transliteration generation // Duan X Y, Banchs R E, Zhang M, et al. Proceedings of the Fifth Named Entity Workshop. Beijing: ACL, 2015: 61–66

[4] Koehn P, Och F J, Marcu D. Statistical phrase-based translation // Daelemans W, Osborne M. Proceedings of the HLT/NAACL. Edmondon: ACL, 2003: 127–133

[5] Stolcke A. SRILM — an extensible language modeling toolkit // Hansen J H L, Pellom B L. Proceedings of International Conference on Spoken Language Processing. Denver: Interspeech, 2002: 901–904

[6] Chen S F, Goodman J T. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98. Cambridge: Computer Science Group, Harvard University, 1998

[7] Och F J, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19–51

[8] Koehn P, Hoang H, Birch A, et al. Moses: open source toolkit for statistical machine translation // Carroll J A, Van den Bosch A, Zaenen A. Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demonstration Session. Prague: ACL, 2007: 177–180

[9] Li H Z, Ma B, Lee C H. A vector space modeling approach to spoken language identification. IEEE Transaction on Acoustic, Speech, Signal Processing, 2007, 15(1): 271–284

[10] Liu L, Finch A, Utiyama M, et al. Agreement on targetbidirectional LSTMs for sequence to sequence learning // Schuurmans D, Wellman M P. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016: 2630–2637

[11] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate // CoRR. Ithaca, New York, 2014: abs/1409.0473

[12] Chen B, Zhang M, Li H, et al. A comparative study of hypothesis alignment and its improvement for machine translation system combination // Su K Y, Su J, Wiebe J. Proceedings of ACL-IJCNLP. Singapore: ACL, 2009: 941–948

[13] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation // Proceeding of AMTA. Boston: AMTA, 2006: 223–231

[14] Melamed I D. Models of translational equivalence among words. Computational Linguistics, 2000, 26 (2): 221–249

[15] He X, Yang M, Gao J, et al. Indirect HMM-based hypothesis alignment for combining outputs from machine translation systems // Lapata M, Ng H T. Proceedings of EMNLP. Hawaii: ACL, 2008: 98–107

[16] Zeiler M D. ADADELTA: an adaptive learning rate method // CoRR. Ithaca, New York, 2012: abs/1212.5701

[17] Cheng Y, Shen S, He Z J, et al. Agreement-based joint training for bidirectional attention-based neural machine translation // Kambhampati S. Proceedings of IJCAI. New York: AAAI, 2016: 2761–2767